Metron manuscript No.
(will be inserted by the editor)

A new Bayesian approach for determining the number of components in a finite mixturee

METRON DOI:10.1007/s40300-015-0068-1

Murray Aitkin · Duy Vu · Brian Francis

Received: 10 October 2014 / Accepted: 17 June 2015

Acknowledgements We are grateful for research support from the Australian Research Council under project DP120102902 for the support of Duy Vu for the period of this research (20122015), and for visits by Brian Francis from the University of Lancaster.

Murray Aitkin
University of Melbourne, Vistoria, Australia. E-mail: murray.aitkin@unimelb.edu.au

Duy Vu
University of Melbourne, Vistoria, Australia. E-mail: duy.vu@unimelb.edu.au

Brian Francis
Lancaster University, Lancaster, UK. E-mail: B.Francis@Lancaster.ac.uk

A new Bayesian approach for determining the number of components in a finite mixture

June 16, 2015

Abstract

This article evaluates a new Bayesian approach to determining the number of components in a finite mixture. We evaluate through simulation studies mixtures of normals and latent class mixtures of Bernoulli responses. For normal mixtures we use a "gold standard" set of population models based on a well-known "testbed" data set – the galaxy recession velocity data set of Roeder (1990). For Bernoulli latent class mixtures we consider models for psychiatric diagnosis (Berkhof, van Mechelen and Gelman 2003).

The new approach is based on comparing models with different numbers of components through their *posterior deviance distributions*, based on non-informative or diffuse priors.

Simulations show that even large numbers of closely spaced normal components can be identified with sufficiently large samples, while for latent classes with Bernoulli responses identification is more complex, though it again improves with increasing sample size.

1 Background: the number of components problem

Finite mixture models are now in widespread use: McLachlan and Peel (2000) give a detailed and authoritative review. Computational methods for maximum likelihood and Bayesian analyses through the EM algorithm and Markov chain Monte Carlo analyses are routinely used and are well-documented. An outstanding remaining problem is the number of components which can be identified in a finite mixture. Chapter 6 of McLachlan and Peel discusses this at length and reports some simulation results for mixtures of multivariate normals with moderately large samples. A more recent study by Nylund, Asparouhov and Muthen (2007) considers latent class models and growth mixture models, with simulations again from moderate to large samples. An issue not addressed in these simulations is the performance of procedures in small samples, where power may be low, especially for latent class models with Bernoulli responses.

Aitkin (2001) reviewed at length Bayesian analyses of mixtures of normals for the galaxy data of Roeder (1990) and found wide variations in the number of components identified, through comparisons of their integrated likelihoods, by different analysts. These analyses required the specification of (hyper) parameters used in the proper priors for the integrated likelihoods. Different specifications of these parameters, and/or the priors, led to different conclusions about the number of components. No "default" analysis with non- or minimally-informative priors was possible. No simulations were reported for any of the analyses, leaving open the performance of these and any other procedures.

Aitkin (1997, 2010) extended Dempster's (1997) treatment, of the posterior distribution of the likelihood for testing simple null hypotheses, to the comparison of arbitrary models, proposed a new Bayesian analysis based on the comparison of deviance distributions under each model, and applied it to the galaxy data in Aitkin (2010, 2011). Again no simulations were reported, so the difference in his conclusions from those based on integrated likelihoods did not lead to a clarification.

The present paper investigates the properties of the deviance distribution approach to Bayesian model comparisons in finite mixtures. We first set out in Section 2 the properties of the frequentist and Bayesian analyses of finite mixtures, illustrated by mixtures of normals and latent class mixtures of Bernoullis. We describe the inferential difficulties of both frequentist and current Bayesian methods for determining the number of components in the mixture. We then give the new approach based on the posterior distribution of the deviance.

Section 3 discusses the galaxy data which illustrates the normal mixture problem, and provides a simulation study of the performance of the deviance distribution approach on galaxy-like data sets, compared with the DIC approach of Spiegelhalter et al (2002).

Section 4 discusses the psychiatric symptom data which illustrates the latent class problem, and provides a simulation study of the performance of the deviance distribution approach on similar data sets.

Section 5 gives conclusions.

2 Models and methods

2.1 The normal mixture model

The general model for a K-component normal mixture for a response variable Y has different means μ_k and variances σ_k^2 in each component:

$$f(y) = \sum_{k=1}^{K} \pi_k f(y|\mu_k, \sigma_k)$$

where

$$f(y|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{1}{2\sigma_k^2}(y-\mu_k)^2\right\}$$

and the σ_k and π_k are positive with $\sum_{k=1}^{K} \pi_k = 1$.

Given a sample $y_1, ..., y_n$ from $f(y)$, the likelihood is $L(\theta) = \prod_{i=1}^{n} f(y_i)$, where $\theta = (\pi_1, ..., \pi_{K-1}, \mu_1, ..., \mu_K, \sigma_1, ..., \sigma_K)$.

An important practical question is why we should assume that the component distributions are normal. Efficient computation of maximum likelihood estimates and posterior distributions can be achieved for a wide range of continuous or discrete mixture distributions, not restricted to the exponential family.

One plausible reason is that mixtures of normal distributions can reproduce a very wide range of distributional shapes. However if our substantive interest

is in the component densities, then it *does* matter whether we use a mixture of normal or, for example, t or lognormal distributions, which may need a different number of components. We return to this question in the discussion of the galaxy data in §3.

2.2 Bernoulli mixtures

We deal also with Bernoulli response data, and use the example discussed at length in §4, of psychiatric patients with a number of symptoms of psychiatric illness. We define $y_{ij} = 1$ if patient i has symptom j, and $y_{ij} = 0$ otherwise, and write p_{ij} for the probability that patient i has symptom j. A simple unstructured model for all of the the y_{ij} for patient i would be a product Bernoulli model:

$$\Pr[\{y_{ij}\} \mid i] = \prod_{j=1}^{r} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}},$$

in which symptoms are possessed independently within a patient. Assuming independence also of the y_{ij} across patients, the likelihood of the observed data would then be

$$\Pr[\{y_{ij}\}] = \prod_{i=1}^{n} \prod_{j=1}^{r} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}.$$

A simple specific model for the table is the *Rasch* model, in which

$$\operatorname{logit} p_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = \theta_i + \phi_j,$$

where θ_i is the *patient propensity* to have any symptom, and ϕ_j is the *symptom propensity* for any patient. A simpler model is the Rasch "symptom only" model, which omits the θ_i parameters. The assumption of complete independence of symptom possession within a patient however appears unreasonable, especially as we aim to identify subgroups of symptoms which tend to occur together within classes, so are not independent.

2.2.1 The latent class model

We assume there are distinct sets of symptoms for (unobservable) classes of psychiatric illness. These are specified by a latent class model, in which the probability of symptom j is q_{jk}, constant for patients in latent class k, but varying among classes, and we assume the weaker conditional independence of symptoms within classes. If the proportion of the psychiatric patient population

in class k is π_k, the probability of the observed data is given by

$$\begin{aligned}
\Pr[\{y_{ij}\} \mid k, i] &= \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1-y_{ij}} \\
\Pr[\{y_{ij}\} \mid i] &= \sum_{k=1}^{K} \left[\pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1-y_{ij}} \right] \\
\Pr[\{y_{ij}\}] &= \prod_{1=1}^{n} \left\{ \sum_{k=1}^{K} \left[\pi_k \prod_{j=1}^{r} q_{jk}^{y_{ij}} (1 - q_{jk})^{1-y_{ij}} \right] \right\}.
\end{aligned}$$

2.3 Fitting the model

The EM algorithm is the standard frequentist tool for fitting finite mixtures, and is discussed at length in McLachlan and Peel (2000), who give the galaxy data as one of their examples (pp 104-5, 194-6). Unobserved indicator variables Z_{ik}, which are 1 if observation i belongs to component k, and zero otherwise, provide the "missing data" aspect of the EM algorithm. In the E step of the algorithm the unobserved indicators in the complete data log-likelihood are replaced by their conditional expectations given the observed data and the current parameter estimates; in the M step the expected complete data log-likelihood is maximized to give new parameter estimates.

A major difficulty with mixture models is the occurrence of local maxima of the likelihood in addition to the global maximum. The usual advice for this problem (for example McLachlan and Peel 2000 p. 55) is to use many random starting points for the EM algorithm, either as random parameter values for the first E step or probabilistic assignments of observations to components for the first M step. The number of random starting points might be set at 100, 1000, 5000 or more – we examine these choices for the galaxy data in §3. As the number of model components increases, we should increase the number of random starting points as there are approximately n^K possible random assignments of observations to components, and a fixed number of starting points will sample more and more sparsely from the possible configurations as K increases. We can expect the number of local maxima to increase as well, for example from slightly different posterior probabilities of assignment of observations to components.

If a local maximum is close in likelihood to the global maximum, but has different assignments of observations to components, it will clearly not be possible to interpret the component assignment posterior probabilities from the global maximum as soundly based; these are in any case usually based on plug-in ML estimates for the model parameters, and so they overstate the precision of the probability of component assignment.

Bayesian "fitting" is also straightforward using the DA (Data Augmentation) algorithm (Tanner and Wong 1987), a special case of Markov Chain Monte Carlo

in which the unobserved indicators and the parameters are drawn alternately: the indicators from their conditional distribution given the observed data and the current parameter draws, and the parameters from their conditional distribution given the observed data and the current indicator draws.

2.4 The number of components – frequentist methods

As is well known, the likelihood ratio test statistic does not have the usual asymptotic χ^2 distribution for nested models when comparing models with different numbers of components. Alternative frequentist decision criteria include AIC and BIC, and bootstrapping the likelihood ratio test statistic from a sequence of fitted models of components. Frequentist and Bayesian analyses are conceptually straightforward using the EM algorithm and Gibbs sampling respectively, provided that for the normal mixture model the component standard deviations are bounded below in some way to prevent single or multiple identical observations defining components with variances converging to zero. Such "singleton" components are outside the model specification (since we specify $\sigma_k > 0$) and may represent recording errors. They are irrelevant in the context of clumping of galaxies – a clump for a single galaxy does not add to our understanding of galaxy clustering.

An alternative is to bound the ratio of largest to smallest component variance. This approach and its generalisation to mixtures of multivariate normals is discussed in Garcia-Escudero, Gordaliza, Matran and Mayo-Iscar (2015).

For the Bayesian analysis, local modes which are far away (in likelihood) from the global model can be ignored, since the posterior probability of the parameter set for such a mode is very low.

2.5 The number of components – Bayesian methods

Bayesian methods for determining the number of components are conventionally based on the integrated likelihoods $\bar{L}_k(\phi_k) = \int L_k(\theta_k)\pi_k(\theta_k \mid \phi_k)\mathrm{d}\theta_k$ for each model k, integrated with respect to the prior distribution $\pi_k(\theta_k \mid \phi_k)$ of the unspecified model parameters θ_k with specified ϕ_k. The ratio of two such integrated likelihoods is called the Bayes factor for their comparison, and is interpreted as though it were the likelihood ratio for two *completely specified* models (Kass and Raftery 1995). However the integrated likelihood is not uniquely, or even conventionally, defined by the likelihood, as it depends on the specification of the prior $\pi_k(\theta_k \mid \phi_k)$ and its parameters ϕ_k – conventional improper diffuse priors cannot be used.

This can lead to very different integrated likelihoods for different proper prior specifications for the same model and data, and these differences are inherent in the definition of the integrated likelhoods, and do not disappear with increasing sample size (Aitkin 2001). An alternative (Dempster 1997, Aitkin 1997, 2010) is to use the *posterior distributions* of the likelihoods, by substituting M (typically 10,000) random draws $\theta_k^{[m]}$ of the parameters θ_k from their posterior into the

likelihoods $L_k(\theta_k)$, giving M corresponding random draws $L_k^{[m]} = L_k(\theta_k^{[m]})$ from the posterior distributions of the likelihoods.

Because of the scale of likelihoods, as in the frequentist analysis we use deviances $D_k(\theta_k) = -2 \log L_k(\theta_k)$ rather than likelihoods L. Models are then compared for the *stochastic ordering* of their posterior deviance distributions, initially by graphing the cdfs of the deviance draws $D_k^{[m]} = D_k(\theta_k^{[m]})$ for each number of components.

The DIC of Spiegelhalter et al (2002) also uses these deviance draws, but only to compute the *mean deviance* across the draws. The DIC, like AIC and BIC and some other decision criteria, requires a *penalty* (in this case using the *effective number of parameters*) on the mean deviance to account for model complexity.

This is not needed for the comparison of deviance distributions: models with increasing numbers of components are effectively penalized for their increasing parametrization, as they have increasingly diffuse deviance distributions because of the decreasing data information about each component. The practical use of deviance distributions is illustrated in the following sections.

3 The galaxy data

The galaxy data published by Roeder (1990) are the recession velocities, in units of 10^3 km/sec, of 82 galaxies from six well-separated conic sections of space; the tabled velocities are said by astronomers to be in error by less than 0.05 units. Roeder noted that the distribution of velocities is important, as it bears on the question of "clumping" of galaxies: if galaxies are clumped by gravitational attraction, the distribution of velocities would be multi-modal; conversely, if there is no clumping effect, the distribution would increase initially and then gradually tail off.

We do not analyse separately the data from the six regions, following all authors including the astronomers Postman et al. (1986) who gave the full data by region. The individual regions have very small data sets, from which not much can be learned about clustering among or within regions. The data are reproduced below, ordered from smallest to largest and scaled by a factor of 1000 as in Roeder (1990) and Richardson and Green (1997). The empirical cdf of the velocities is shown in Figure 1, together with the fitted normal distribution cdf.

```
Recession velocities (/1000) of 82 galaxies

 9.172  9.350  9.483  9.558  9.775 10.227 10.406 16.084
16.170 18.419 18.552 18.600 18.927 19.052 19.070 19.330
19.343 19.343 19.440 19.473 19.529 19.541 19.547 19.663
19.846 19.856 19.863 19.914 19.918 19.973 19.989 20.166
20.175 20.179 20.196 20.215 20.221 20.415 20.629 20.795
20.821 20.846 20.875 20.986 21.137 21.492 21.701 21.814
21.921 21.960 22.185 22.209 22.242 22.249 22.314 22.374
22.495 22.746 22.747 22.888 22.914 23.206 23.241 23.263
23.484 23.538 23.542 23.666 23.706 23.711 24.129 24.285
24.289 24.366 24.717 24.990 25.633 26.960 26.995 32.065
32.789 34.279
```

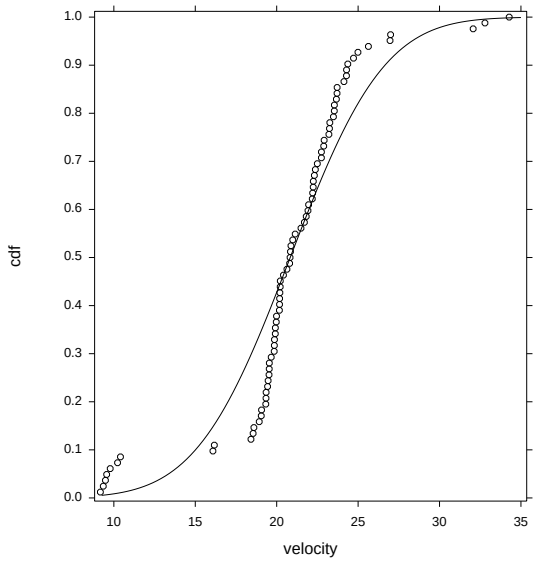

Figure 1: Empirical cdf (o) and fitted normal cdf (–) for the galaxy data

It is immediately clear that the normal distribution does not fit, with a gap or jump between the seven smallest observations around 10 and the large central body of observations between 16 and 26, and another gap between 27 and 32, for the three largest observations. Following many other authors, we assume that the mixture of normal distributions is appropriate: we give some support for this below.

Maximum likelihood estimates for the galaxy data can be found in Aitkin (2001, 2010), and in many other references. Using the probit vertical scale clarifies the improvement in fit with increasing numbers of components, from 1 to 4, shown in Figures 2–5.

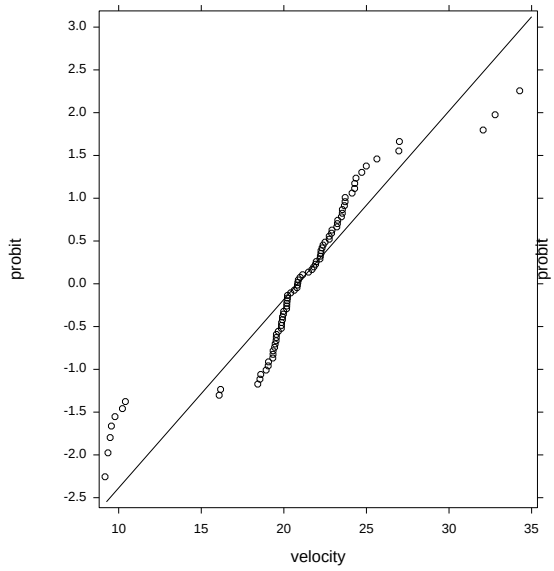

Figure 2: K=1

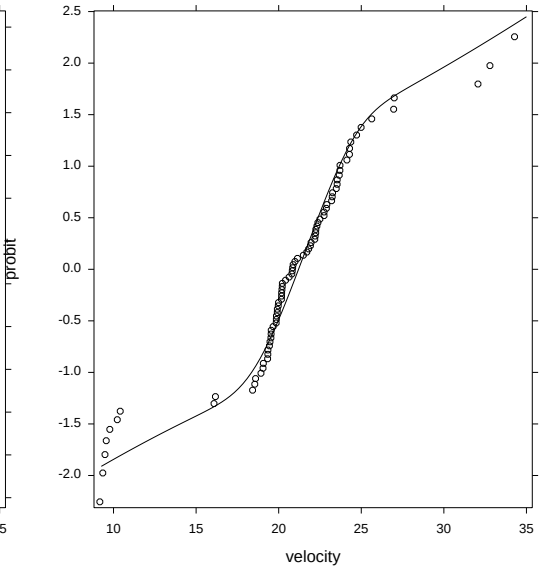

Figure 3: K=2

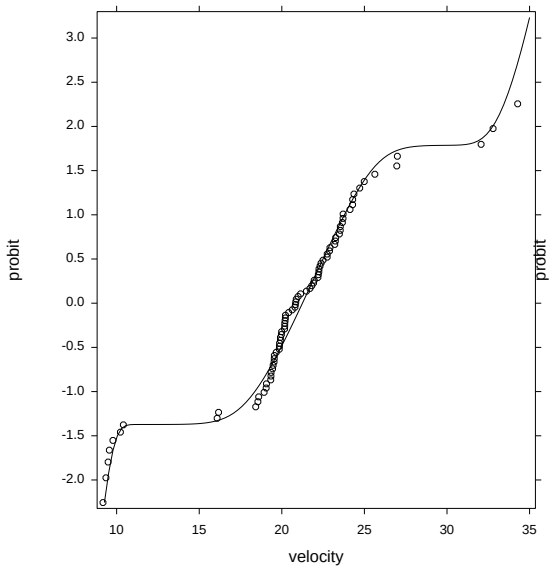

Figure 4: K=3

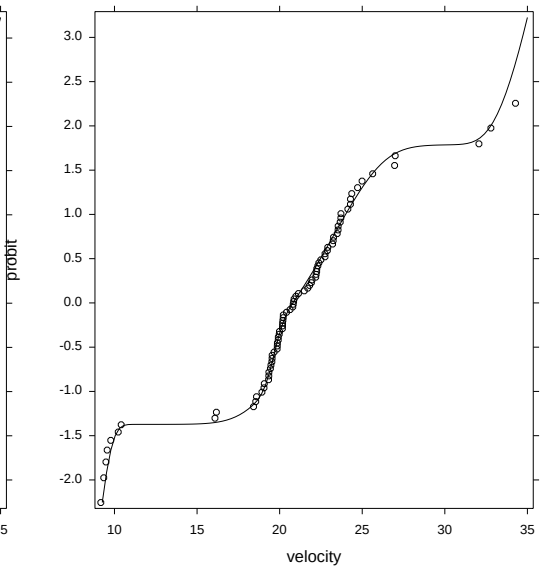

Figure 5: K=4

The two-component model does not give a good fit to the extreme observations on both sides. The three-component model gives a close fit except in the velocity interval 18-22, where it seems to have slightly the wrong slope. The four-component model gives a very close fit, with slightly different standard deviations for the two central components (0.45 and 2.27). The probit scale shows that the component velocity distributions are nearly linear on this scale, supporting the normal distribution assumption within component.

3.1 How many components?

3.1.1 Frequentist methods

Frequentist deviances, AICs and BICs are given for each number of components up to 6 in Table 1 (adapted from Aitkin 2010). The number of model parameters is $p = 3K - 1$.

K	p	deviance	AIC	BIC
1	2	480.83	482.83	485.24
2	5	413.78	423.78	435.82
3	8	406.96	422.96	437.83
4	11	395.43	417.43	443.94
5	14	392.27	420.27	454.01
6	17	365.15	399.15	440.12

Table 1: Deviance, AIC and BIC for K components

AIC selects $K = 6$ and BIC selects $K = 2$. The bootstrap likelihood ratio test was used by McLachlan and Peel (2000, p. 196) for the galaxy data in 100 bootstrap replications for $K = 1, ..., 6$, and gave bootstrap p-values, for testing K components against $K + 1$, of 0.01, 0.01, 0.01, 0.04, 0.02 and 0.22, suggesting $K = 6$.

The BIC choice of $K = 2$ does not seem well-supported by the cdf plot, with departures in both tails. Given the close fit of the observed data cdf to that for the 3- and 4-component mixtures, where does the evidence for six components come from in the AIC/bootstrap conclusions? This is provided by the two sets of closely-paired observations (16.084, 16.170) and (26.960, 26.995), which define components with extremely small variances, giving a large reduction in deviance. The 3- and 4-component cdf plots show that these points are very close to the fitted cdf in both plots – there is little evidence of the need for two additional small components.

There seem to be some conflicts in the frequentist conclusions, both among methods and with the cdf plots.

3.1.2 Bayesian methods

The galaxy data have been analysed many times by Bayesian methods, mostly using the integrated likelihood. Detailed discussions of these analyses can be

found in Aitkin (2001, 2010, 2011). The DA algorithm was used in most analyses, with proper priors on the component means, variances and proportions. The integrated likelihoods were then converted to posterior model probabilities through Bayes's theorem, with either a flat prior distribution on the number of components or an informative proper prior (a truncated Poisson distribution was a common choice). It is difficult to compare the posteriors for the number of components when the priors are different.

A quite different form of posterior analysis – RJMCMC (reversible jump MCMC) – was used by Richardson and Green (1997) in which the number of components was included as a discrete parameter in the joint parameter space, and MCMC analysis included this parameter, "jumping" across the different parameter spaces for different numbers of components. This gave a direct computation of the posterior for the number of components, without requiring the computation of integrated likelihoods, but at the cost of very heavy and complex MCMC computations.

Here we reproduce from Aitkin (2011) the posterior distributions for the number of components using the analysts' priors, and their conversion to the equivalent posteriors for a flat prior. For the rescaled analyses by Escobar and West, Phillips and Smith, and Stephens, the posterior probabilities for extreme values of K could not be computed from the limited precision given in the available results, and are represented by question marks. All these posterior distributions were decreasing beyond the last value given, and we assume they continue to do so with increasing K. The unknown tail values have been ignored in rescaling the posteriors to sum to 1. The initials refer to:

- EW: Escobar and West (1995)
- PS: Phillips and Smith (1996)
- S: Stephens (2000)
- RW: Roeder and Wasserman (1997)
- RG: Richardson and Green (1997).

K	1	2	3	4	5	6	7	8	9	10
EW	.01	.06	.14	.21	.21	.17	.11	.06	.02	
PS	.16	.24	.24	.18	.10	.05	.02	.01		
S	.58	.29	.10	.02	.004	.001	-	-	-	
RW	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10
RG	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03...

Table 2: Prior distributions for K

K	3	4	5	6	7	8	9	10	11	12	13
EW	-	.03	.11	.22	.26	.20	.11	.05	.02	-	-
PS	-	-	-	.03	.39	.32	.22	.04	-	-	-
S	.55	.34	.09	.01	-	-	-	-	-	-	-
RW	.999	.00	-	-	-	-	-	-	-	-	-
RG	.06	.13	.18	.20	.16	.11	.07	.04	.02	.01	.01

Table 3: Posterior distributions for K

K	3	4	5	6	7	8	9	10	11	12	13
EW		.01	.03	.07	.13	.18	.30	.28	?	?	?
PS				.00	.10	.21	.43	.26	?	?	?
S	.10	.25	.35	.29	?	?	?	?	?	?	?
RW	$>.999$	$<.001$									
RG	.06	.13	.18	.20	.16	.11	.07	.04	.02	.01	.01

Table 4: Posterior distributions for K (flat prior)

It is immediately striking that the posteriors for the EW and PS analyses have modes at $K = 9$, with high probability also for $K = 10$, while that for RG has its mode at 6, is very diffuse, and does not rule out $K = 9$, or $K = 3$ or 4. The posterior for S has a mode at $K = 5$ and a slightly lower value at $K = 6$. The RW distribution is almost a spike at $K = 3$. The PS posterior rules out $K \leq 6$.

It is hard to imagine a more diverse, and inconsistent, set of posterior conclusions about a parameter across these five papers. In the discussion of Aitkin (2001), and in Stephens (2000), this difference is obscured by the strongly informative priors for K used by EW, PS and S, which almost eliminate the possibility of $K \geq 9$. If we have no prior view about the number of mixture components, which conclusions are believable?

There is an obvious difficulty with the EW and PS conclusions, and that is the sample size relative to the number of model parameters. With seven components in the mixture, the *average* sample size per parameter is only four. For $K \geq 9$, the two extreme groups are further split into subgroups with single observations, for which standard deviations cannot be estimated – the ML analysis of the model breaks down at this point. It seems therefore unbelievable that the Bayesian analysis can give nine or 10 components with high probability.

Aitkin (2001, 2011) gave a discussion of the priors and their specified parameter values used in these analyses. Here we refer only to the point made above: that when integrated likelihoods are used for Bayesian model comparisons, their values depend explicitly on both the priors used and the settings of their parameters. So no conclusions can be drawn from these analyses about the number of components in the mixture.

3.1.3 Posterior deviance analysis

For the posterior deviance analysis by the Data Augmenation algorithm in a computational framework like BUGS (Gibbs sampling in this application), we need proper priors. These are proper but diffuse: a diffuse Dirichlet prior (with indices 1,...,1) on the component proportions π_k, diffuse conjugate priors on the means μ_k – normal priors with zero means and variances 100 – and diffuse, again with large variances, conjugate priors on the inverse variances σ_k^{-2}.

The galaxy data were analysed in this framework by Celeux *et al* (2006). We summarise[1] their analysis. After convergence of the Gibbs sampler, 10,000 values were sampled from the thinned posterior distributions and the K-component mixture deviances computed for each parameter set. These were kindly supplied by Gilles Celeux.

We show in Figure 6 the deviance distributions for $K = 1, ..., 7$ on the same scale (more detailed Figures are given in Aitkin 2011).

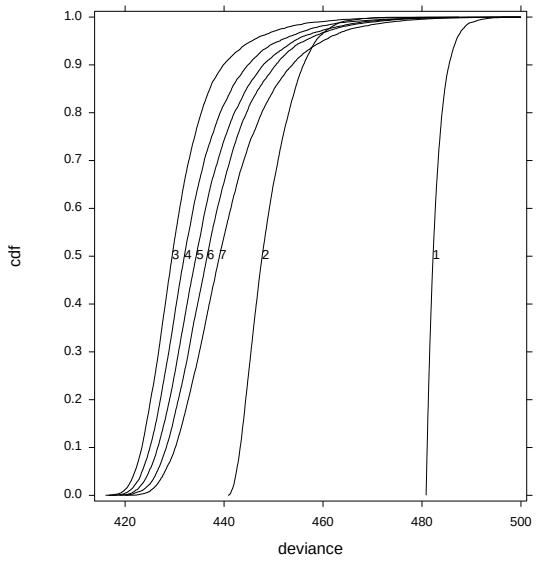

Figure 6: Deviances for 1-7 components

The interpretation of this figure can be simply summarised:

- The deviance distribution for $K = 2$ greatly improves on that for the single normal $K = 1$.
- The improvement continues for $K = 3$.
- As the number of components increases from 3 to 7 the deviance distributions move steadily to the right, to larger deviance values (lower likelihoods).

[1] Full details can be found in their paper.

- They also become more *diffuse*, with increasing slope.

3.1.4 Stochastic ordering

A real random variable X is *less than* another real random variable Y *in stochastic order* if

$$\Pr(X > a) \leq \Pr(Y > a) \ \forall a \in (-\infty, \infty),$$

with a strict inequality for at least one a. If so, we say that X is *stochastically strictly less than* Y, and X and Y are *stochastically ordered*. This is equivalent to the inequality $F_X(a) \geq F_Y(a)$, with the cdf of X strictly greater than that of Y at some point a. (We need to allow for the fact that the cdfs will be identical at $\pm\infty$.)

We apply this concept to the model deviance distributions.

> If the cdf of the deviance distribution for model k is stochastically less than that of model k', we say that model k *fits the data better* than model k', or that model k *is preferred to* model k'.

The extent of the preference is assessed by the posterior distribution of the probability of model k. If this preference is only slight, then we are unable to choose confidently between the models.

So the deviance distributions for $K = 3$ to 7 are *stochastically ordered*, with $K = 3$ *fitting the data best* (of the normal mixture distributions): these distributions do not cross because of their steady movement to the right and the increasing slopes of the cdfs with increasing parametrization. All these distributions are stochastically ordered with respect to that for $K = 1$. The distribution for $K = 2$ is not stochastically ordered with respect to those for $K = 4 - 7$ as it crosses them. The distributions for $K = 2$ and 3 converge at about the 99th percentile. We conclude from the stochastic orderings that the best model has $K = 3$.

This conclusion agrees with that of Roeder and Wasserman, and the Celeux et al analyses, but is inconsistent with the other Bayesian analyses. Since *none* of the existing Bayesian analyses have been supported by simulations, we investigate how well our criteria perform in simulations from galaxy-like data sets generated from known mixture models. We note first however that in this example the *membership* of each galaxy in the three (or more) components is not an issue of particular interest, and we do not comment on it here. In the next example however this is a very important issue, and we discuss it in detail there.

It may happen (and does in simulations) that the deviance distributions for the competing models are not stochastically ordered because the best two or more deviance distributions cross. In this case we need to consider more carefully the comparison of the deviance draws. At the m-th deviance draw from each model we have deviances $D_k^{[m]}$. For each k and m we define indicator variables $W_k^{[m]} = 1$ if $D_k^{[m]}$ is the smallest of the K deviances, $W_k^{[m]} = 0$ otherwise. We

aggregate over the draws, to give:

$$W_k^+ = \sum_{m=1}^{M} W_k^{[m]}.$$

We assessed two possible criteria for "best model":

- the model with the lowest *median* deviance (this is similar to one of the versions of the DIC with the median deviance replacing the mean deviance, but without the penalty);
- the model with the largest W_k^+ – the model "most often best" across the draws.

3.2 Simulation studies

In the first study, we generated data sets of size $n = 82$ from normal mixture distributions with $K = 1$ to 7 components, with 100 data sets from each, with parameters given by the ML estimates from the galaxy data (given in Aitkin 2010 p. 213) with the corresponding number of mixture components. The number of observations from each component was conditioned to give each component its mixture proportion multipled by the sample size and rounded, so that the structure of the galaxy data set was closely approximated.

For each generated data set, we fitted (Bayesianly) from 1-7 normal mixture components, and obtained the posterior distribution of the deviance for each number of components. These seven deviance distributions were then compared, and the best chosen by the smallest median deviance and the most often best criteria. The DIC was also computed for each K, and the value of K with the smallest DIC was taken as the "best" in the DIC comparison framework.

For each K, we give in Table 5 the percentage of correct identifications in the 100 data sets by both the DIC and the most often best ("mob") criterion (the smallest median deviance criterion was consistently inferior to the most often best, and is not reported). Model identification was very successful for small K, but fell off dramatically beyond $K = 3$.

True K	DIC	mob
1	100	100
2	85	98
3	51	99
4	3	9
5	0	18
6	2	9
7	0	1

Table 5: Model identification $n = 82$

In the second study, we successively doubled the 100 sample sizes to $n = 164$, 328 and 656, with parameters as before. Table 6 gives the percentages of correct identification using the DIC and the most often best criterion.

n	82		164		328		656	
K	DIC	mob	DIC	mob	DIC	mob	DIC	mob
1	100	100	100	99	100	100	100	100
2	85	98	100	100	100	100	100	97
3	51	99	98	99	100	99	100	99
4	3	9	11	67	30	99	17	99
5	0	18	0	9	0	37	1	89
6	2	9	0	10	56	100	78	100
7	0	1	0	15	4	3	4	32

Table 6: Percentages of correct model identification in 100 data sets of size n using DIC and most often best posterior deviance

The deviance distribution criterion was consistently more accurate than the DIC, which had difficulty with more than three components in all the sample sizes considered. As n increased, so did the successful identification of the correct model. With a sufficiently large sample, even the 6-component model could be correctly identified by the posterior deviance. The 7-component model required a larger sample size.

The occasional non-monotonicity of the identification proportion with both K and n is due partly to the structure of the galaxy data (for example the components are better separated in the 6-component than in the 5-component model) and partly to the small simulation size.

4 The psychiatric symptom data

The data set used by Berkhof, van Mechelen and Gelman (BMG) came from a previous study by van Mechelen and De Boeck (1989), which assessed the presence or absence of $r = 23$ possible binary symptoms in $n = 30$ psychiatric patients. Their approach aimed to assess the identification of different classes of psychiatric illness in subgoups of patients. The data are shown in Table 7 (where x indicates symptom present, . indicates symptom absent).

1	disorientation	.....x..
2	obsession/compulsion	...x.....
3	memory impairment	.....x.....x..
4	lack of emotion	.....x.....x
5	antisocial impulses or acts	....x.....xx.....
6	speech disorganization	.....x...x.x
7	overt anger	.....x.....x.....x....
8	grandiosity	.....x..x.....x..x.....
9	drug abuse	x...x.....x.....x...
10	alcohol abuse	.....x.....xx.....x.x..
11	retardation	.....x..xx....x.x
12	belligerence/negativism	.....x.....xxx.....x...
13	somatic concerns	..x.....xx.....x.....x.xx
14	suspicion/ideas of persecution	.....xxx.....xx...xx
15	hallucinations/delusions	.....xxx.....xx...xx
16	agitation/excitement	.....x.....x.xx.....xxx..x.
17	suicide	.x....xxx..xx...xx...xx.....xx
18	anxiety	.xxx..xxxxxx...x.x..xxx.xxx..x
19	social isolation	x.....xx.xxxx..xx.xxxx.x.xxxx
20	inappropriate affect or behaviour	...xx.x..x....xxxx.xxxxxxxxxx
21	depression	xxx...xxxxxxxx..xx.xxxx..xx.xx
22	leisure time impairment	..xxxxxxxxxxxx.xxxxxxxxxxx.x
23	daily routine impairment	..xxxxxxxxxxxx.xxxxxxxxxxxx

Table 7 - symptom data in 30 patients

We define $y_{ij} = 1$ if patient i has symptom j, and $y_{ij} = 0$ otherwise, and write p_{ij} for the probability that patient i has symptom j. As described in §2.2, we use the latent class model to represent the unobserved classes of psychiatric illness.

4.1 Frequentist analysis of the symptom data

We extend the Rasch model to the latent class models. We summarise in Table 8 the frequentist deviances, number of model parameters, AIC and BIC for the null, Rasch and latent class models with up to four classes.

Model	deviance	# params	AIC	BIC
Null	844.68	1	846.68	848.08
Rasch	571.96	52	675.96	748.82
$K = 1$	606.54	23	652.54	684.74
$K = 2$	534.30	47	628.30	694.10
$K = 3$	473.04	71	615.04	714.44
$K = 4$	439.03	95	629.03	762.03

Table 8: Frequentist deviances, symptom data

AIC chooses the 3-class model, BIC the 1-class model. The deviance changes from 1-2, 2-3 and 3-4 classes are respectively 72.21, 61.26 and 34.01, all with 24 degrees of freedom, and corresponding naive p-values of $10^{-6}, 4 \times 10^{-4}$ and .084. It appears from the likelihood ratio test (naively interpreted) that three classes are necessary.

4.2 Bayesian analysis

4.2.1 Priors

BMG used a diffuse Dirichlet (1,...,1) prior for the class mixture probabilities π_k, and independent Beta(α, α) priors for the class-specific symptom probabilities q_{jk}. They ran MCMC to obtain posterior distributions for the model parameters θ (the sets of π_k and q_{jk}) and the class membership indicators Z_{ik} in the complete data representation used for MCMC:

$$\Pr[\{y_{ij}\} \mid \{Z_{ik}\}] = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{r} \left[\pi_k q_{jk}^{y_{ij}} (1 - q_{jk})^{(1-y_{ij})} \right]^{Z_{ik}}.$$

4.2.2 Model comparison through integrated likelihoods

BMG compared models with 1,...,5 classes through their integrated likelihoods. They set the α parameter of the common Beta priors for the q_{jk} equal to 0.5, 1 and 2, and compared the models with $K = 1,..,5$ classes at each α. When the models were compared at

- $\alpha = 2$ (an informative quadratic prior), there was a preference for the one class model;
- $\alpha = 1$ (the uniform prior), there was equal preference for the two- and three-class models;
- $\alpha = 0.5$ (the Jeffreys prior), there was a preference for the three-class model.

It is clear that the preferred number of classes (in terms of the largest integrated likelihood) is a direct function of α, with the preferred number of classes increasing with decreasing α. Since both the uniform and the Jeffreys priors are

widely used as "reference" or "minimally informative" priors, this form of prior specification does not lead to a clear preference for the number of classes.

4.2.3 Varying the prior

In an expanded sensitivity analysis, BMG changed the common prior for the symptom probabilities q_{jk} to Beta(α, β), with α and β being determined by the data in an empirical Bayes approach. They used a diffuse hyperprior density for (α, β), uniform on $\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{(\alpha+\beta)^{\frac{1}{2}}}\right)$, in the range

$$\frac{\alpha}{\alpha+\beta} \in (0,1),\ \frac{1}{(\alpha+\beta)^{\frac{1}{2}}} \in (0,c), c>0,$$

and then estimated α and β from the posterior mode $\bar{\alpha}, \bar{\beta}$. The log integrated likelihoods using this approach for K = 1,...,5 classes were -346.9, -340.8, -335.8, -335.7 and -335.8. So three, four and five classes were almost equally well-supported, with very weak support for two classes and no support for one class. They conjectured that the very small differences in the integrated likelihoods for three, four and five classes were because the number of patients was too small to be able to draw a distinction between them.

BMG discussed the further need to determine whether the priors are *reasonable for these data* in terms of the prior predictive distribution, by generating random parameter draws from the priors, and generating random data sets from the models, given the values of the model parameters.

They compared properties of the simulated data sets with those of the real data set, to assess which priors were consistent with the data. They concluded that the symmetric Beta(α, α) priors were *not* consistent with the data, while the asymmetric Beta$(\bar{\alpha}, \bar{\beta})$ prior *was* consistent.

So BMG's conclusion, after considerable further effort which we do not give here, was that the number of classes was probably three.

4.3 The role of the prior

A serious concern for the analyst following this approach is the need to *check the prior against the data* – since the conclusions are strongly affected by variations in the priors, it seems obvious that the priors themselves should be checked for reasonableness – they become part of the model structure.

This however conflicts with a fundamental principle of Bayesian analysis: that priors are specified *before the data are observed* – they should not be *tuned to the data* after observing them. Updating the prior based on the observed data – the likelihood – gives the *posterior*, not a "reasonable" prior consistent with the data. The need for this "consistency" arises only because of the integration of the likelihood over the prior used for the comparison of models.

4.4 Bayesian model comparison through posterior deviances

We now apply the posterior deviance analysis to the psychiatric symptom data. Models with different numbers of latent classes are compared in the same way as for the galaxy data.

With diffuse or reference priors $\pi(\theta_k)$ on the model parameters θ_k for model k, we make M independent draws $\theta_k^{[m]}$ from the parameter posterior $\pi(\theta_k \mid \mathbf{y})$, substitute them into the model deviance $D_k(\theta_k) = -2 \log L_k(\theta_k)$, and compare across k the cumulative distributions of the deviances $D_k^{[m]} = D_k(\theta_k^{[m]})$ for stochastic ordering.

Three other models act as *reference* models for the latent class model:

- The *null* model, with a single common symptom frequency parameter for all patients;
- the *Rasch* model, an additive model in patient and symptom on the logit scale; it is the simplest logit model reproducing the row and column marginal totals of the data array;
- the *saturated* model has a different set of Bernoulli symptom parameters for each patient – each patient is a "separate class".

The posterior deviance distributions for the latent class models from $M = 10,000$ draws with up to three classes, and for the Rasch and saturated models, are shown in Figure 7. The saturated model deviance is very diffuse, and is well to the right of those for the latent class models. It has so many parameters that each is very poorly defined. (The null model deviance, not shown, is much the worst: it increases from its minimum of 844.68 to 850, off the scale of the Figure.) Figure 8 shows, on a larger scale, the distributions for one to five classes. In interpreting the Figures, the *leftmost* distribution, if it does not cross any another, is *stochastically smallest* and identifies the preferred model. If two deviance distributions cross, the strength of preference for one over the other is determined by the percentile at which they cross.

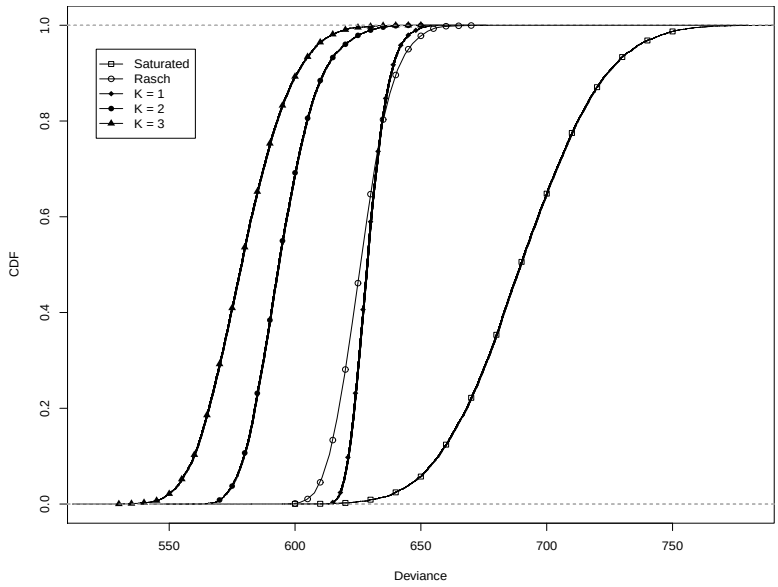

Figure 7: Psychiatric symptom deviance distributions, K=1-3, Rasch and saturated models

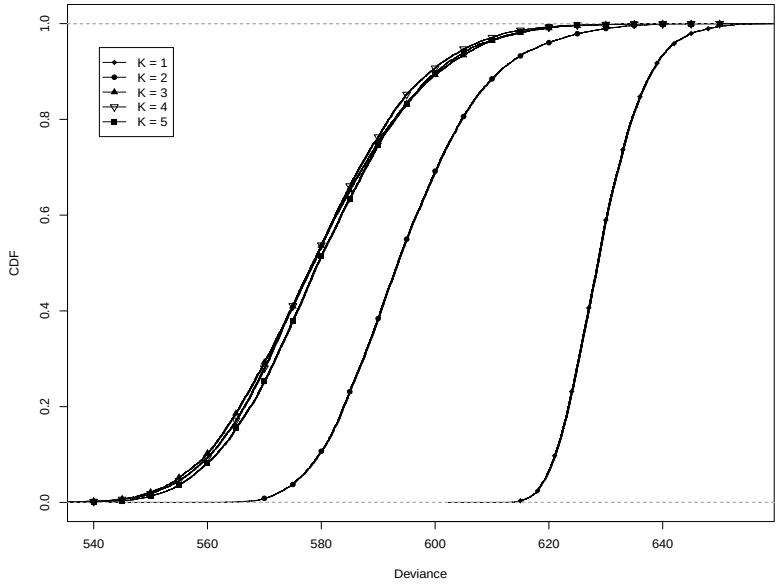

Figure 8: Psychiatric symptom deviance distributions, K=1-5

4.5 Conclusions from the posterior deviance model comparison

The results from the posterior deviance cdfs displayed in Figures 7 and 8 can be summarized as follows:

- The saturated model deviance is the worst of those shown (each Bernoulli parameter has just one observation, of 0 or 1).[2]
- The Rasch model is a poor fit – the latent class distributions improve on it substantially.
- The class deviance distributions shift substantially to the left from $K = 1$ to 2, and by a further 10 (at the median) from $K = 2$ to 3.
- The distributions for $K = 3$, 4 and 5 overlap very closely, and intersect.
- For K increasing beyond 5, the distributions move slowly to the right (not shown, but they are computed for up to 15 classes).

So *three classes are clearly identified*, but the evidence for more than three is confusing. Figure 8 explains the strange similarity of the 3-, 4- and 5-class integrated likelihoods: the spacing between these deviance distributions is so small that a one-point integrated summary gives very close integrated likelihoods.

Without clear stochastic ordering for the 3-, 4- and 5-class models, we compared the models for K = 1-5 through the median and most often best criteria as for the galaxy example. The 4-class model was the best by both criteria.

The comparison of median deviances shows the same features as the integrated likelihoods. We compare them in Table 9, converting the integrated likelihoods $\bar{L}_k$ to the deviance scale of $-2 \log \bar{L}_k$:

K	1	2	3	4	5
median dev.	628.57	593.44	578.66	578.50	579.43
int.-lik.-dev.	693.8	681.6	671.6	671.4	671.6

Table 9: Median and integrated-likelihood-based deviances for the symptom data

The deviances from the integrated likelihoods are substantially larger than the median deviances, and are off the scale of Figure 8. They correspond to parameter sets with very low likelihood. This is a well-known feature of integrated likelihoods, which tend to zero with increasing diffuseness of the priors, leading to *Lindley's paradox* (Bartlett 1957, Lindley 1957, Kass and Raftery 1995). The slow deterioration of model fit with increasing K and the extent of

[2]This may seem counter-intuitive, since the *ML estimate* of the saturated model always has the smallest *frequentist* deviance. The single observation in each "class" however gives a very diffuse likelihood for each p_{ij} and this leads to a very diffuse and large deviance distribution.

overlap appear to be characteristic of Bernoulli latent class models, and to be more severe than for mixtures of normals.

Because of the small sample size of patients, and the absence of psychiatric opinion on the existence and relevance of sub-classes of patients, we do not discuss these in detail here. However, though three classes appear to be established, one of these is very small, and even the 2-class model gives class differences on only four of the symptoms.

4.6 Simulation studies

In the first study, we generated 100 data sets of size $n = 30$ from from latent class models with $K = 1,...,5$ classes, with parameters given by the MLEs from the symptom data. For each generated data set, we fitted (Bayesianly) 1-5 classes, and obtained the posterior distribution of the deviance for each number of classes. The five deviance distributions were then compared, and the"best" chosen by the smallest median deviance and by the most often best (mob) criteria (Table 10). The DIC was not used in this part of the study.

True K	median	mob
1	79	76
2	79	80
3	58	55
4	27	33
5	19	21

Table 10: Model identification n = 30

Four or five classes are unlikely to be identified in the sample size we have and with population parameters equal to the sample MLEs, and the chance of identifying three classes is only around 50%. In the second study, we successively doubled the sample size to $n = 60$ and 120, with parameters as before (duplicating the patient parameters). Table 11 gives the percentages of correct identification using the mob criterion.

$K \backslash n$	30	60	120
1	76	100	100
2	80	91	100
3	55	100	99
4	33	82	100
5	21	15	51

Table 11: Model identification n = 30, 60, 120

Again with a larger sample of 120 we are able to identify correctly 4 classes, but the 5-class model appears to require more data.

5 Conclusion

The comparison of competing mixture models through their posterior deviance distributions worked successfully in simulations, given a sufficiently large sample. Unsurprisingly, more classes have more parameters and need larger samples.

The three components identified for the galaxy data are obvious to the eye, and are as much as could be expected from the sample size and the pattern of generating parameter values. The small sample size and overlap of the deviance distributions for the symptom data make inference about the number of classes particularly difficult.

It would be a question for the psychiatrist whose patients form the data set to comment on whether these differences do establish a clear sub-class of symptoms identifying distinct sub-categories of psychiatric illness.

The focus on categories of illness is a consequence of working with the latent class model, but it is not the only possible model: a latent variable model with a normal "propensity to have symptoms" could also be considered. The Rasch model considered above is one such, which gives equal weight to each symptom; it gave a much worse deviance distribution than the 2-class model. Weighting the symptoms differentially leads to the 2PL (two-parameter logit) model widely used in psychometrics. We investigated this model also: its deviance distribution was between those of the 2-class and 3-class models. It therefore seems possible that for these 30 patients the model of categories of psychiatric illness is inappropriate, and the varying frequencies of symptoms could be consistent with a single continuous factor of symptom propensity. Again this is an issue for the patients' psychiatrist, rather than a firm statistical model conclusion.

From a theoretical perspective, the approach to model comparison through deviance distributions has a number of advantages over other model comparison methods:

- It does not require proper informative priors: improper non-informative priors are sufficient, and we recommend their use provided the posteriors are proper.
- Previous work has shown that varying the priors has little effect on the model comparison, if the likelihoods dominate the priors.
- The computation of the deviance distributions even in complex models requires only the standard MCMC output of the posterior distribution of each model's parameters: no integration over the prior is involved.
- The deviance distribution approach performed better than the DIC (in the galaxy simulations): it appears than the DIC loses information in summarising the deviance draws by their mean.

These advantages come with what may be regarded as a disadvantage: of the comparison of deviance *distributions* rather than of *single-number* integrated or penalized likelihoods. However this is more in accord with the general Bayesian

principle that the post-data information about *any* function of the model parameters and the data should be through its posterior distribution.

A natural question about the asymptotic properties of our approach, in identifying the correct model, is not dealt with in this study. Our concern has been to demonstrate the performance of the procedure in the finite (small-to-moderate) samples with which we work. Monte Carlo error affects the probabilities of correct identification, but this is small (from the 1,000 draws) compared to the variability from the 100 samples, and the two are confounded. With increasing sample size the probability of correct identification goes to 1 or very close to it, in both the examples. We feel that this is the appropriate measure of effectiveness. We leave the asymptotic performance for further investigation.

It may seem disappointing also that our study does not compare our approach with the "standard" Bayes factor approach. The reason is simple: *there is no standard Bayes factor approach*: as in the galaxy example, the different specifications of priors and their parameters makes it impossible to define a standard approach, quite apart from the formidable difficulties of the computation of the integrated likelihoods.

References

1. Aitkin, M.: The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood (with discussion). Statistics and Computing **7**, 253-272 (1997)

2. Aitkin, M.: Likelihood and Bayesian analysis of mixtures. Statistical Modelling **1**, 287-304 (2001)

3. Aitkin, M.: Statistical Inference: an Integrated Bayesian/Likelihood Approach. Chapman and Hall/CRC Press, Boca Raton (2010)

4. Aitkin, M.: How many components in a finite mixture? In Mixtures: Estimation and Applications. ed. K.L. Mengersen, C.P. Robert and D.M. Titterington. Wiley, Chichester (2011)

5. Bartlett, M.S.: A comment on D. V. Lindley's statistical paradox. Biometrika **44**, 533-534 (1957)

6. Berkhof, J., van Mechelen, I., Gelman, A.: A Bayesian approach to the selection and testing of mixture models. Statistica Sinica **13**, 423-442 (2003)

7. Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M.: Deviance information criteria for missing data models. Bayesian Analysis **1**, 651-674 (2006)

8. Dempster, A.P.: The direct use of likelihood in significance testing. Statistics and Computing **7**, 247-252 (1997)

9. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**, 577-588 (1995)

10. Garcia-Escudero, L.A., Gordaliza, A., Matran, C., Mayo-Iscar, A.: Avoiding spurious local maximizers in mixture modeling. Statistics and Computing 01/2015; DOI: 10.1007/s11222-014-9455-3 (2015)

11. Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773-795 (1995)

12. Lindley, D.V.: A statistical paradox. Biometrika **44** 187-192 (1957)

13. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

14. van Mechelen, I., De Boeck, P.: Implicit taxonomy in psychiatric diagnosis: A case study. J. Social and Clinical Psychology **8**, 276-287 (1989)

15. Nylund, K.L., Asparouhov, T., Muthen, B.O.: Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Structural Equation Modeling **14**, 535-569 (2007)

16. Phillips, D.B., Smith, A.F.M.: Bayesian model comparison via jump diffusions. in Markov Chain Monte Carlo in Practice, ed. W.R. Gilks, S. Richardson, D.J. Spiegelhalter. Chapman and Hall/CRC Press, Boca Raton (1996)

17. Postman, M., Huchra, J.P., Geller, M.J.: Probes of large-scale structures in the Corona Borealis region. The Astronomical Journal **92**, 1238-1247 (1986)

18. Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. Roy. Stat. Soc. B **59**, 731-792 (1997)

19. Roeder, K.: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. J. Am. Stat. Assoc. **85**, 617-624 (1990)

20. Roeder, K., Wasserman, L.: Practical Bayesian density estimation using mixtures of normals. J. Am. Stat. Assoc. **92**, 894-902 (1997)

21. Spiegelhalter, D.J., Best, N.., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit. J. Roy. Stat. Soc. B **64**, 583-639 (2002)

22. Stephens, M.: Bayesian analysis of mixtures with an unknown number of components - an alternative to reversible jump methods. Ann. Statist. **28**, 40-74 (2000)

23. Tanner, M. and Wong, W.: The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528-550 (1987)